

EXHIBIT 1

[REDACTED]

Joint Chart Regarding Datasets

Exhibit 1 – Parties’ Joint Chart Regarding Datasets

Disputed Request	Defendants’ Response	Plaintiffs’ Proposed Compromise	Defendants’ Proposed Compromise	[Court’s Use]
RFP 2: The Training Data for Imagen, Bard, and Gemini.	<p>Google proposed that Plaintiffs direct their attention to one exemplary text model and one exemplary image model, and remains open to discussing that reasonable compromise. Plaintiffs responded with a counter-proposal that Google “prioritize” the production of training data for every iteration of the Imagen, Gemini, and Bard models. While that approach is extremely costly and unworkable, Google remains willing to meet and confer with opposing counsel regarding a reasonable way to make accessible to Plaintiffs a reasonable scope of training data for a representative set of relevant models. To the extent Plaintiffs believe any further response to this Request is necessary, Google is willing to meet and confer with Plaintiffs</p>	<p>Production of the Ingested Datasets: [REDACTED] FineWeb, FineWeb2, FineWeb-edu, and Common Crawl used to build and train the Models at Issue. <i>See</i> ECF 234 ¶¶ 1, 2, 6, 11, 109, 119, 128, 159, 169.</p>	<p>Google’s proposed compromise was previously accepted by Plaintiffs and approved by the Court: From the available datasets used to train the models at issue, Plaintiffs identified 19 datasets (totaling over 1.5 PB), supported by references to data cards or other documents showing the datasets were used to train at-issue models. Google then set up a remote environment—hosted for months now at Google’s sole expense—that allows Plaintiffs to access, analyze, and manipulate these datasets.</p>	

	on the relevance of that information, including as to class certification.			
RFP 2: The Training Data for Imagen, Bard, and Gemini.	<p>Google proposed that Plaintiffs direct their attention to one exemplary text model and one exemplary image model, and remains open to discussing that reasonable compromise. Plaintiffs responded with a counter-proposal that Google “prioritize” the production of training data for every iteration of the Imagen, Gemini, and Bard models. While that approach is extremely costly and unworkable, Google remains willing to meet and confer with opposing counsel regarding a reasonable way to make accessible to Plaintiffs a reasonable scope of training data for a representative set of relevant models. To the extent Plaintiffs believe any further response to this Request is necessary, Google is willing to meet and confer with Plaintiffs on the relevance of that</p>	<p>Google’s confirmation that The Pile, RedPajama, RedPajama2 were not used to build or train any of the Models at Issue. And if Google responds in the affirmative, production thereof. <i>See</i> ECF 234 ¶¶ 1, 2, 6, 11, 109, 119, 128, 159, 169.</p>	<p>Google’s proposed compromise was previously accepted by Plaintiffs and approved by the Court: From the available datasets used to train the models at issue, Plaintiffs identified 19 datasets (totaling over 1.5 PB), supported by references to data cards or other documents showing the datasets were used to train at-issue models. Google then set up a remote environment—hosted for months now at Google’s sole expense—that allows Plaintiffs to access, analyze, and manipulate these datasets.</p>	

	information, including as to class certification.			
RFP 3: The Training Data for any Imagen, Bard, and Gemini model(s) currently in development or developed during the pendency of this litigation.	Google proposed that Plaintiffs direct their attention to one exemplary text model and one exemplary image model, and remains open to discussing that reasonable compromise. Plaintiffs responded with a counter-proposal that Google “prioritize” the production of training data for every iteration of the Imagen, Gemini, and Bard models. While that approach is extremely costly and unworkable, Google remains willing to meet and confer with opposing counsel regarding a reasonable way to make accessible to Plaintiffs a reasonable scope of training data for a representative set of relevant models. To the extent Plaintiffs believe any further response to this Request is necessary, Google is willing to meet and confer with Plaintiffs on the relevance of that information, including as to class certification.	Production of the Ingested Datasets: [REDACTED] FineWeb, FineWeb2, FineWeb-edu, and Common Crawl used to build and train the Models at Issue. <i>See</i> ECF 234 ¶¶ 1, 2, 6, 11, 109, 119, 128, 159, 169.	Google’s proposed compromise was previously accepted by Plaintiffs and approved by the Court: From the available datasets used to train the models at issue, Plaintiffs identified 19 datasets (totaling over 1.5 PB), supported by references to data cards or other documents showing the datasets were used to train at-issue models. Google then set up a remote environment—hosted for months now at Google’s sole expense—that allows Plaintiffs to access, analyze, and manipulate these datasets.	

<p>RFP 3: The Training Data for any Imagen, Bard, and Gemini model(s) currently in development or developed during the pendency of this litigation.</p>	<p>Google proposed that Plaintiffs direct their attention to one exemplary text model and one exemplary image model, and remains open to discussing that reasonable compromise. Plaintiffs responded with a counter-proposal that Google “prioritize” the production of training data for every iteration of the Imagen, Gemini, and Bard models. While that approach is extremely costly and unworkable, Google remains willing to meet and confer with opposing counsel regarding a reasonable way to make accessible to Plaintiffs a reasonable scope of training data for a representative set of relevant models. To the extent Plaintiffs believe any further response to this Request is necessary, Google is willing to meet and confer with Plaintiffs on the relevance of that information, including as to class certification.</p>	<p>Google’s confirmation that The Pile, RedPajama, RedPajama2 were not used to build or train any of the Models at Issue. And if Google responds in the affirmative, production thereof. <i>See</i> ECF 234 ¶¶ 1, 2, 6, 11, 109, 119, 128, 159, 169.</p>	<p>Google’s proposed compromise was previously accepted by Plaintiffs and approved by the Court: From the available datasets used to train the models at issue, Plaintiffs identified 19 datasets (totaling over 1.5 PB), supported by references to data cards or other documents showing the datasets were used to train at-issue models. Google then set up a remote environment—hosted for months now at Google’s sole expense—that allows Plaintiffs to access, analyze, and manipulate these datasets.</p>	
--	--	--	--	--